

Um modelo híbrido de dicionário para análise ética

**Candidata Doutorado Daniela América da Silva,
Pesquisador Henrique Duarte Borges Louro,
Estudante MSc Gildárcio Sousa Goncalves,
Prof. Dr. Johnny Cardoso Marques,
Prof. Dr. Luiz Alberto Vieira Dias,
Prof. Dr. Adilson Marques da Cunha,
Prof. Dr. Paulo Marcelo Tasinaffo**



**São Paulo, Brasil
Maio 2021**

Sobre o trabalho



**INFORMATION TECHNOLOGY:
NEW GENERATIONS - ITNG 2021**
Las Vegas, NV, USA- April 12-13, 2021

• **2020 Best Student Paper Award**



"A Hybrid Dictionary Model for Ethical Analysis"
Daniela America da Silva
Brazilian Aeronautics Institute of Technology
Brazil

**INFORMATION TECHNOLOGY: NEW GENERATIONS
ITNG 2021**
Online- April 12-13, 2021

*The INFORMATION TECHNOLOGY: NEW GENERATIONS - ITNG 2021
is pleased to present this Certificate of Award to*

Daniela América da Silva

For the Best Student Paper entitled:
A Hybrid Dictionary Model for Ethical Analysis

Shahram Latifi _____ April 13, 2021 _____

Shahram Latifi, General Chair Date

Agenda

- **Breve Introdução Histórica**
- **Método**
- **Considerações - Perigos e Limitações**
- **Tutorial - Banco de Dados e Python**

Introdução

Breve Introdução Histórica

A Web

Opiniões

**Atividades
diárias**



**Oportunidade
para burlar
regras**

**Coisas
inadmissíveis
na sociedade**

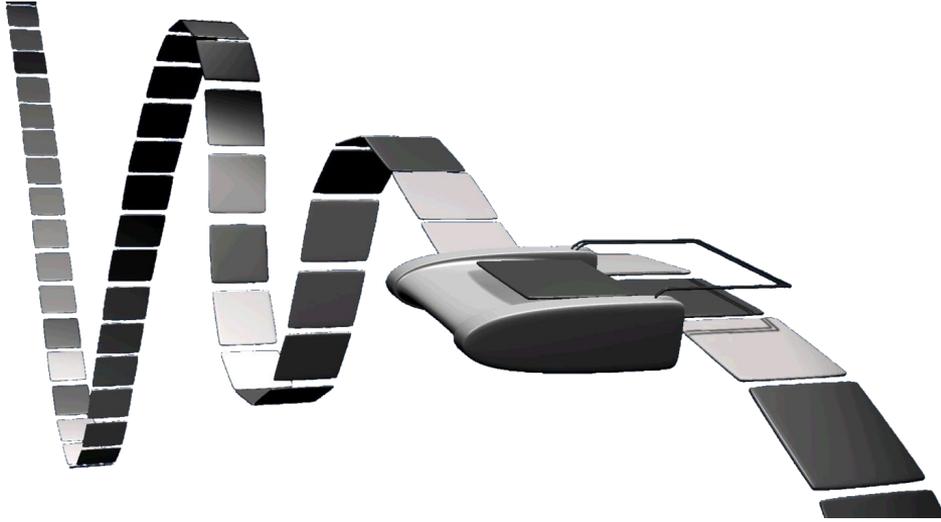
A máquina e o ser humano



A máquina está preparada para o diálogo humano ?

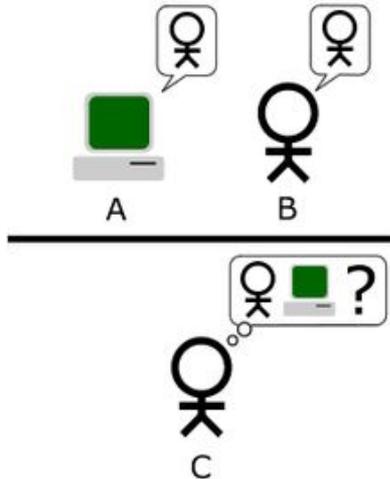
- Abordagem híbrida
- Dicionário
- Reconhecimento de padrões de expressões na Internet.

Allan Turing



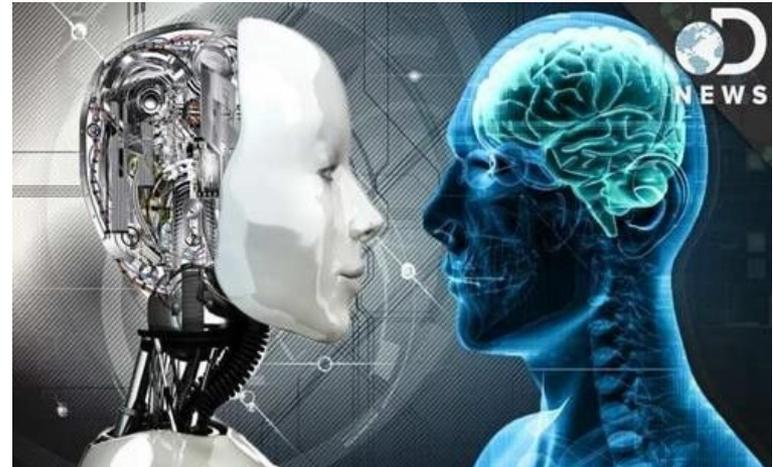
- **Modelo abstrato de um computador, a Máquina de Turing.**
- **Trabalhou na construção de máquinas físicas para quebrar os códigos secretos das comunicações alemãs durante a Segunda Guerra Mundial.**

Teste de Turing



Turing acreditava que a construção desse tipo de máquina dependia apenas de memória, velocidade de processamento e programação adequada.

O Teste de Turing refere-se a quanto uma máquina poderia dialogar com um ser humano, sem que o ser humano neste diálogo fosse capaz de identificar que era uma máquina [2] [3].



A Inteligência Artificial Geral



Como passar no Teste de Turing?

A máquina deve exibir um comportamento semelhante ao comportamento humano denominado Inteligência Artificial Geral.

Desde 1950 a implementação desta IA Geral tem sido pesquisada pela comunidade científica.

O Processamento Natural da Linguagem



References: Web,
[AI Achievements - The Growing Timeline of AI Milestones](https://www.bbc.com/news/technology-39701222)
<https://www.bbc.com/news/technology-39701222>
<https://www.theverge.com/2012/6/26/3119022/alan-turing-60-year-old-chess-program-garry-kasparov>
<https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/>
https://pt.wikipedia.org/wiki/Match_Garry_Kasparov_vs_Deep_Blue



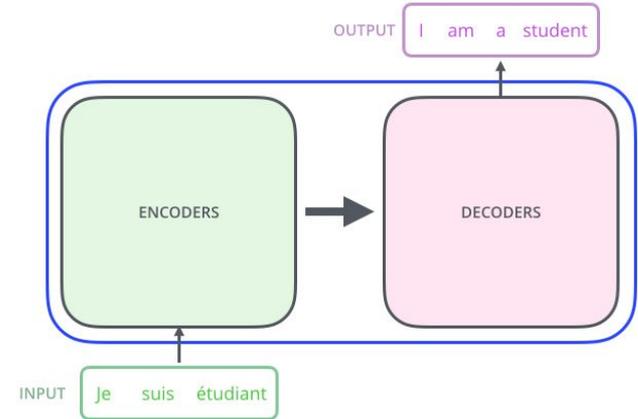
Transformer

O que é o Transformer ?

É uma arquitetura de aprendizagem profunda (deep learning). Utiliza um mecanismo de atenção onde as outras palavras são importantes para o contexto.

Por exemplo: O Maracanã é um palco importante do futebol.

Neste caso, para o mecanismo de atenção, a palavra futebol é mais importante para Maracanã do que a palavra palco.



Método

Método para criação do Dicionário

Método



- **As palavras usadas na vida cotidiana revelam nossos pensamentos, sentimentos, personalidade e motivações.**

Método

- Sugerimos a criação de um dicionário de comportamento antiético, para diversas áreas
- Drogas lícitas; drogas ilícitas; sexo (pedofilia, estupro, abuso, pornografia); armas e armamento; crimes hediondos (roubo, assassinato); contrabando; profanidade; jogos de azar; racismo; homofobia; perjúrio; e difamação.



Método

- Para a construção do dicionário antiético, propomos identificar direto da Internet as palavras encontradas diretamente nos sites com comentários antiéticos e, a seguir, classificar as palavras comumente utilizadas em cada categoria.



Modelo

Parte 1 - Análise de Ocorrências

**Parte 2 - Análise de expressões comuns
para cada categoria**

Parte 3 - Correções Ortográficas

Parte 4 - Política de Remoção de Conteúdo

Parte 1 - Análise de Ocorrências

- A parte 1 do modelo se refere a fornecer uma análise das ocorrências da Internet.
- Nesse modelo, consideramos que os internautas possuem uma forma específica de comunicação.
- Por isso é relevante identificar as palavras e expressões utilizadas por esses usuários para cada categoria de comportamento antiético investigada.



Parte 2 - Análise de expressões comuns para cada categoria

- Por ser um modelo híbrido, a Parte 2 do modelo se refere a fornecer uma análise das expressões comuns para cada categoria.
- Primeiramente, usando como referência o *LIWC Linguistic Inquiry and Word Count Program*, inicialmente desenvolvido pela Universidade do Texas com a Universidade de Auckland [4], o dicionário deve ter duas características centrais: uma para processamento, que abre uma série de comentários da Internet; e a outra para verificar palavra por palavra e a partir do dicionário, o que é aplicável ao comportamento antiético.
- Em segundo lugar, utilizando abordagem semelhante à adotada pela Universidade Federal de Minas Gerais (UFMG, Brasil), no artigo *A Measurement Study of Hate Speech in Social Media* [5]. buscando identificar por categoria uma estrutura de expressão em comentários da Internet.

Parte 3 - Correções Ortográficas



- A parte 3 do modelo refere-se à correções ortográficas. Atualmente, existem vários corretores ortográficos disponíveis e usados dinamicamente.
- No caso de análise de comportamento antiético, o corretor ortográfico será usado estaticamente uma vez que os comentários já foram registrados.
- o objetivo é corrigir palavras e conjugações de verbos, de modo que o comentário seja apropriado para aplicação do dicionário e para encontrar o padrão de comunicação em uma categoria específica.

Parte 4 - Política de Remoção de Conteúdo

- A parte 4 do modelo refere-se ao desenvolvimento de uma política de remoção de conteúdo. O objetivo é evitar que algumas remoções sejam interpretadas como discriminatórias, por isso é importante que esta política seja divulgada no Portal dos utilizadores.



REMOVE

content that violates our policy as quickly as possible



RAISE

up authoritative voices when people are looking for breaking news and information



REWARD

trusted, eligible creators and artists



REDUCE

the spread of content that brushes right up against our policy line

Exemplo de política de remoção de conteúdo do Youtube

Exemplo dicionário em Inglês - LIWC

Area	Some examples
I. STANDARD LINGUISTIC DIMENSIONS	Word Count, Words per sentence, Sentences ending with ? question marks, pronouns (I, our), 1st person singular (I, my), 1st person plural (we, us), total first person (I, we), second person (you), third person (she, them), negations (no, never), assents (yes, OK, mmhmm), articles (a, the), prepositions (on, to), numbers (one, thirty)
II. PSYCHOLOGICAL PROCESSES	Affective or emotional processes (happy, ugly), positive emotions (happy, pretty), positive feelings (happy, love), optimism and energy (pride, win), negative emotions (hate, enemy), anxiety or fear (nervous, tense), anger (hate, pissed), sadness or depression (grief, sad), cognitive processes (cause, know), causation (because, effect), insight (think, know), discrepancy (should, would), inhibition (block, constrain), tentative (maybe, perhaps), certainty (always, never), sensory and perceptual (see, touch), seeing (view, look), hearing (heard, listen), feeling (touch, hold), social (talk, us), communication (talk, share), friends (pal, buddy), family (mom, cousin), humans (boy, woman)



Exemplo dicionário em Inglês - LIWC

III. RELATIVITY	Time (hour, day), past tense verb (walked, were), present tense verb (walk, is), future tense verb (will, might), space (around, over), up (up, above), down (down, under), inclusive (with, and), exclusive (but, except), motion (walk, move)
IV. PERSONAL CONCERNS	Occupation (work, class), school (class, student), job or work (employ, boss), achievement (try, goal), leisure (house, TV), home (house, kitchen), sports (football, game), television and movies (TV, sitcom), music (tunes, song), money and financial issues (cash, taxes), metaphysical issues (God, heaven), religion (God, church), death and dying (dead, burial), physical states and functions (ache, breast), body states (ache, heart), sex and sexuality (lust), eating, drinking, dieting (eat, taste), sleeping, dreaming (bed, dreams), grooming (wash, bath)
APPENDIX: EXPERIMENTAL DIMENSIONS	Swear words (damn, piss), nonfluencies (uh, rr), fillers (you know, I mean)



Exemplo de estudo de expressões em Inglês - UFMG

| < intensity > < userintent > < hatetarget >

List of words used as <intensity> token: *absolute, absolutely, actually, already, also, always, bloody, completely, definitely, do, especially, extremely, f"cking, fckin, fkn, fr, freakin, freaking, fucken, fuckin, fucking, fuckn, generally, genuinely, honestly, honesty, jus, just, kinda, legitimately, literally, naturally, normally, now, officially, only, passively, personally, proper, really, realy, rly, rly, secretly, seriously, simply, sincerely, so, sometimes, sorta, srsly, still, strongly, totally, truly, usually*

Twitter	% posts	Whisper	% posts
I hate	70.5	I hate	66.4
I can't stand	7.7	I don't like	9.1
I don't like	7.2	I can't stand	7.4
I really hate	4.9	I really hate	3.1
I fucking hate	1.8	I fucking hate	3.0
I'm sick of	0.8	I'm sick of	1.4
I cannot stand	0.7	I'm so sick of	1.0
I fuckin hate	0.6	I just hate	0.9
I just hate	0.6	I really don't like	0.8
I'm so sick of	0.6	I secretly hate	0.7

Table 1: Top ten hate intent in Twitter and Whisper.

List of synonyms of "hate": *do not like, abhor, despise, detest, loathe, scorn, shun, abominate, anathematize, contempt, curse, deprecate, deride, disapprove, disdain, disfavor, disparage, execrate, nauseate, spurn, am allergic to, am disgusted with, am hostile to, am loath, am reluctant, am repelled by, am sick of, bear a grudge against, cannot stand, down on, feel malice to, have an aversion to, have enough of, have no use for, look down on, do not care for, object to, recoil from, shudder at, spit upon*

Twitter		Whisper	
Hate target	% posts	Hate target	% posts
Nigga	31.11	Black people	10.10
White people	9.76	Fake people	9.77
Fake people	5.07	Fat people	8.46
Black people	4.91	Stupid people	7.84
Stupid people	2.62	Gay people	7.06
Rude people	2.60	White people	5.62
Negative people	2.53	Racist people	3.35
Ignorant people	2.13	Ignorant people	3.10
Nigger	1.84	Rude people	2.45
Ungrateful people	1.80	Old people	2.18

Table 2: Top ten targets of hate in Twitter and Whisper.

List of words to exclude from the first hate word pattern: *about, all, any, asking, disappointing, everyone, following, for, having, hearing, how, hurting, is, it, letting, liking, many, meeting, more, most, my, myself, on, other, seeing, sexting, some, telling, texting, that, the, them, these, this, those, watching, wen, what, when, when, whenever, why, with, you*

Exemplo de Expressões Bullying - Dimensões

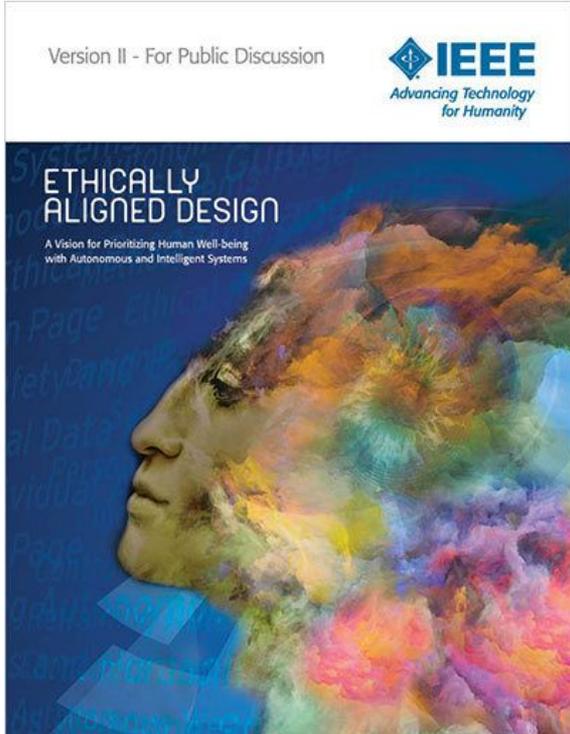
Tipo	Variável	Vocabulário
Intensidade	Ênfase	absoluto, absolutamente, na verdade, já, também, sempre, sangrenta, completamente, definitivamente, faça, especialmente, extremamente, porra, geralmente, genuinamente, honestamente, honestidade, jus, apenas, legitimamente, literalmente, naturalmente, normalmente, agora, oficialmente, somente, passivamente, pessoalmente, adequado, realmente, real, secreto, seriamente, simplesmente, sinceramente, então, às vezes, meio que, ainda, fortemente, totalmente, verdadeiramente, geralmente
	Insulto	execrar, nausear, sou alérgico a, estou com nojo, sou hostil a, sou repugnante, sou relutante,
Intenção	Humilhação	não gosto, abomino, desprezo, detesto, desprezar, amaldiçoar, depreciar, ridicularizar, desaprovar, desdém, desfavor, depreciação,
	Violência psicológica	sou repellido por, estou cansado de, guardar rancor contra, não posso ficar de pé, sentir malícia, ter aversão a, ter o suficiente
	Intimidação	não use, olhe para baixo, não ligue, faça objeções, recue, estremecer,
	Constrangimento	cuspir



Considerações

Perigos e Limitações

Desenvolvimento alinhado com a ética



Transparência: a base de uma decisão de uma sistema autônomo e inteligente (A/SI) particular deve ser sempre detectável;

Responsabilidade: A/SI deve ser criado e operado de acordo com uma lógica inequívoca para todas as decisões tomadas;

Conscientização do uso indevido: os criadores de A/SI devem se proteger contra todos os possíveis usos indevidos e riscos do A/SI em operação.

Os perigos e limitações

nature View all journals Search Login

Content Journal info Publish Sign up for alerts RSS feed

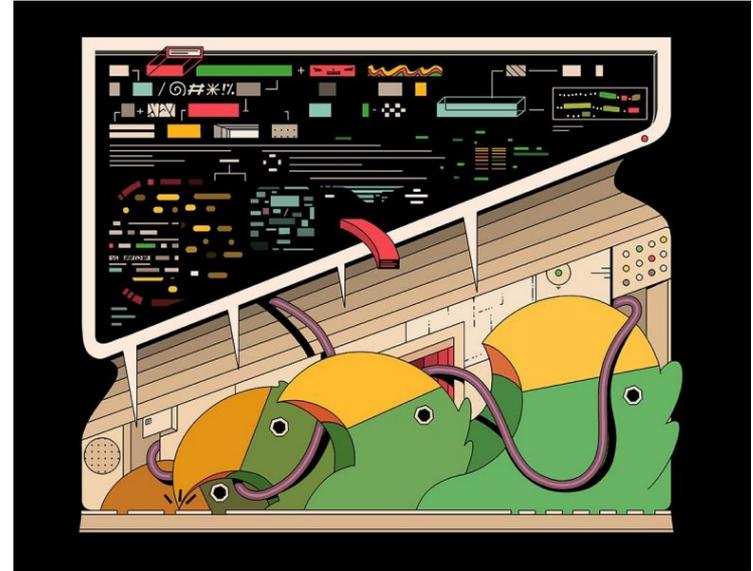
nature > news feature > article

NEWS FEATURE · 03 MARCH 2021

Robo-writers: the rise and risks of language-generating AI

A remarkable AI can write like humans – but with no understanding of what it's saying.

Matthew Hutson



Os perigos e limitações

Kevin Lacker's blog About Me

Giving GPT-3 a Turing Test

Jul 6, 2020

I've been playing around with OpenAI's new [GPT-3](#) language model. When I got beta access, the first thing I wondered was, how human is GPT-3? How close is it to passing a Turing test?

How It Works

Let me explain how exactly I'm generating these conversations. GPT-3 is a general language model, trained on a large amount of uncategorized text from the internet. It isn't specific to a conversational format, and it isn't trained to answer any specific type of question. The only thing it does is, given some text, guess what text comes next.

So if we want GPT-3 to generate answers to questions, we need to seed it with a "prompt". I'm using this prompt to initialize all of the Q&A sessions:

Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?

Q: How many eyes does the sun have?

A: The sun has one eye.

Q: How many eyes does a blade of grass have?

A: A blade of grass has one eye.

Q: How do you sporgle a morgle?

A: You sporgle a morgle by using a sporgle.

Q: How many bonks are in a quoit?

A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?

A: It takes two rainbows to jump from Hawaii to seventeen.

Q: Which colorless green ideas sleep furiously?

A: Ideas that are colorless, green, and sleep furiously are the ideas of a sleep furiously.

Os perigos e limitações

Full Citation: Nguyen A, Yosinski J, Clune J. *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*. In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

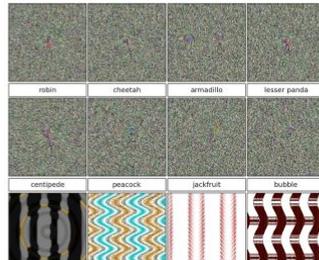
Anh Nguyen
University of Wyoming
anguyen8@uwyo.edu

Jason Yosinski
Cornell University
yosinski@cs.cornell.edu

Jeff Clune
University of Wyoming
jeffclune@uwyo.edu

Abstract

Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are now able to classify objects in images with near-human-level performance, questions naturally arise as to what differences remain between computer and human vision. A recent study [30] revealed that changing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). Here we show a related result: it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99%



- 1 State-of-the-art DNNs can recognize real images with high confidence
- 2 But DNNs are also easily fooled: images can be produced that are unrecognizable to humans, but DNNs believe with 99.99% certainty are natural objects

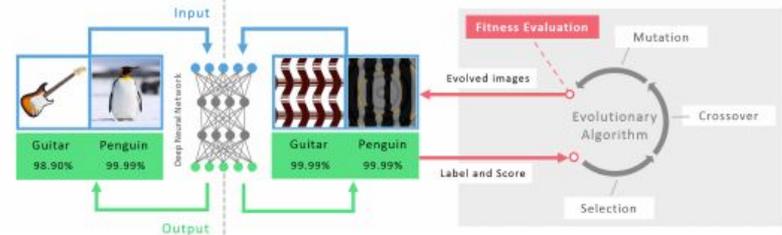


Figure 2. Although state-of-the-art deep neural networks can increasingly recognize natural images (left panel), they also are easily fooled into declaring with near-certainty that unrecognizable images are familiar objects (center). Images that fool DNNs are produced by evolutionary algorithms (right panel) that optimize images to generate high-confidence DNN predictions for each class in the dataset the DNN is trained on (here, ImageNet).

Banco de dados e Conceitos de PNL em Python

Demonstração do Banco de Dados

Contar Palavras do Dicionário	<pre>select count(*) from palavra;</pre>
Contar Palavras ligadas à comentários antiéticos por categoria	<pre>select a.tip_id, min(tip_descricao), count(*) from palavra_has_tipo a, tipo b where a.tip_id = b.tip_id group by a.tip_id;</pre>
Contar Comentários capturados do Twitter	<pre>select count(*) from ocorrencia;</pre>
Identificar Período de Coleta de Dados	<pre>select min(ocr_criacao), max(ocr_criacao) from ocorrencia;</pre>
Listar as 30 palavras mais relevantes conforme número de ocorrências	<pre>select b.pal_palavra , a.qtd from (select tho.tip_id tip_id, pt.pal_id pal_id, count(*) qtd from tipo_has_ocorrencia tho, palavra_has_tipo pt, tipo t where tho.tip_id = t.tip_id and pt.tip_id = t.tip_id group by tho.tip_id, pt.pal_id) a, palavra b where a.pal_id = b.pal_id order by 2 desc;</pre>

Demonstração de Conceitos de PNL (Python)

Em tempos de incerteza, a educação se revela ainda mais necessária e urgente. E a pandemia do coronavírus tem sido implacável ao expor em muitos de nós a falta de informação e o despreparo para lidar com temas relacionados ao uso seguro da internet, sobretudo por crianças e adolescentes.

Longe fisicamente da escola e dos amigos, o distanciamento social é compensado por um intensa vida online, marcada pelo aumento repentino no tempo de uso de aparelhos conectados à internet. A COVID-19 está transformando também a internet. Mas será que estamos preparados para lidar com os riscos e mediar as interações dessa vida ultraconectada? Será que as crianças e adolescentes têm ferramentas para fazerem um uso positivo e crítico das redes?

Referências:

<https://new.safernet.org.br/content/como-pandemia-vem-reforcando-importancia-da-educacao-digital>

Demonstração de Conceitos de PNL (Python)

Tamanho do texto

Tokenize: separar o texto por sentenças e por palavras

Relevância: as palavras mais utilizadas

Pontuação: eliminar caracteres de pontuação ou especiais

Stopwords: eliminar as palavras que podem não ser representativas no texto

Steaming: identificar uma palavra e suas múltiplas formas

Part of Speech e Lemmatization: identificar se uma palavra é verbo, pronome, advérbio, substantivo

Demonstração de Conceitos de PNL (Python)

https://colab.research.google.com/github/dasamerica/tutorials/blob/main/NLP/ExemploPortugues_towardsAI_natural_language_processing_tutorial.ipynb

The screenshot displays a Google Colab notebook titled "ExemploPortugues_towardsAI_natural-language-processing-tutorial.ipynb". The interface includes a top navigation bar with options like "Arquivo", "Editar", "Ver", "Inserir", "Ambiente de execução", "Ferramentas", and "Ajuda". A left sidebar shows a table of contents with sections such as "Adaptado de TowardsAI por Daniela América com exemplos em português", "Próximos passos disponíveis apenas para o Inglês", "Bag of Words", and "Seção". The main content area features a title "Adaptado de TowardsAI por Daniela América com exemplos em português" and "Natural Language Processing (NLP) with Python-Tutorial". Below the title, there are two bullet points: "Tutorial: <https://colab.research.google.com/drive/1vt4u8Hp-iQIRSFc6330utvl7t3R34-IJ?usp=sharing#scrollTo=8CbkOmvBVXPY>" and "Github: https://github.com/towardsai/tutorials/tree/master/natural_language_processing". The code cell contains the following text:

```
#To get the text file used :  
#!wget https://raw.githubusercontent.com/towardsai/tutorials/master/natural_language_processing/Natural_Language_Processing_Text.txt  
  
!wget https://raw.githubusercontent.com/dasamerica/tutorials/master/NLP/Natural_Language_Processing_Text.txt  
  
#To get the image file used :  
#!wget https://raw.githubusercontent.com/towardsai/tutorials/master/natural_language_processing/circle.png  
  
!wget https://raw.githubusercontent.com/dasamerica/tutorials/master/NLP/circle.png
```

The bottom of the notebook shows a taskbar with several tabs, including "ExemploPortugu...ipynb" (Verified) and "circle.png" (Verified). A button labeled "Exibir todos" is visible on the right side of the taskbar.

Conclusão

- Este artigo teve como objetivo apresentar um modelo de dicionário híbrido para análise de sentimento aplicada a comportamentos antiéticos.
- Quanto ao aprendizado de máquina, não existe um modelo matemático para o comportamento humano. E o poder de processamento não é suficiente para afirmar que a máquina tem uma inteligência semelhante à humana. É necessário um modelo para a criação de dicionários que auxiliam no processamento de comentários do usuário na Internet.
- Este trabalho focou especificamente na identificação de categorias de comportamento antiético na Internet, criando o dicionário da Internet por meio de um comitê, identificando padrões em expressões e analisando políticas de remoção de conteúdo.
- Esta investigação mostrou que a colaboração humana é necessária para avaliar o contexto do comentário e julgar quais palavras podem ser mantidas, removidas e adicionadas a um dicionário. Recomenda-se uma equipe multidisciplinar para julgar quais palavras são relevantes para o dicionário.

Agradecimentos

- Os autores agradecem pelo apoio geral e financeiro durante o desenvolvimento deste Projeto Acadêmico:

- O Instituto Tecnológico de Aeronáutica (ITA)



- A Empresa Ecosystema Negócios Digitais Ltda

- A Fundação Casimiro Montenegro Filho (FCMF)



- E pelo apoio geral para esta apresentação:

- A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)



- A Comunidade de Estatística do Prof. Thiago e a EstatiDados



Referências

- Lamdgrebe J, Smith B. There is no general AI: Why Turing machines cannot pass the Turing test;2019.
- Turing Test. 2019;Available from: <https://en.wikipedia.org/wiki/Turingtest>
- Teste de Turing. 2019;Available from:<https://pt.wikipedia.org/wiki/TestedeTuring> .
- Tausczik Y, Pennebaker J. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods;2010.
- Mondal M, et all. A Measurement Study of Hate Speech in Social Media;2017.
- da Silva D.A. et al. (2020) A Hybrid Dictionary Model for Ethical Analysis. In: Latifi S. (eds) 17th International Conference on Information Technology–New Generations (ITNG 2020). Advances in Intelligent Systems and Computing, vol 1134. Springer, Cham.
https://doi.org/10.1007/978-3-030-43020-7_83
- Shukla, et al., “Natural Language Processing (NLP) with Python — Tutorial”, Towards AI, 2020

Contato

- **Silva, Daniela América, Candidata ao Doutorado**
- **das.america@gmail.com**
- **Instituto Tecnológico de Aeronáutica**
- **Programa de Engenharia Eletrônica e Computação - Área: Informática**

